The Genealogy of Ideology: Predicting Agreement and Persuasive Memes in the U.S. Courts of Appeals

Shivam Verma*
Department of Mathematics
Courant Institute of Mathematical Sciences
New York University
sv1239@nyu.edu

Adithya Parthasarathy*
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University
ap4608@nyu.edu

Daniel L. Chen
Institute for Advanced Study
Toulouse School of Economics
daniel.li.chen@gmail.com

Abstract

We employ machine learning techniques to identify common characteristics and features from cases in the US courts of appeals that contribute in determining dissent. Our models were able to predict vote alignment with an average F1 score of 73%, and our results show that the length of the opinion, the number of citations in the opinion, and voting valence, are all key factors in determining dissent. These results indicate that certain high level characteristics of a case can be used to predict dissent. We also explore the influence of dissent using seating patterns of judges, and our results show that raw counts of how often two judges sit together plays a role in dissent. In addition to the dissents, we analyze the notion of memetic phrases occurring in opinions - phrases that see a small spark of popularity but eventually die out in usage - and try to correlate them to dissent.

1 Introduction

Past and recent advances in machine learning techniques and Natural Language Processing (NLP) augur an increase in their use and importance in the analysis of legal literature. A number of recent studies use machine learning on Supreme Court and other law-related datasets to make interesting predictions, such as predicting the outcome of Supreme Court decisions [1], something which legal experts are notoriously unsuccessful at [2], or predicting authorship of unsigned judicial opinions [3].

The opportunity and challenge in prediction and inference problems in the field of law lies in the underlying text form of the datasets.

Our overarching objective is to predict how two judges in a particular panel align on their voting, based on the historical vote alignment of that judge with other judges. In addition to simply using the voting history, we also attempt to make use of the citation history among cases as well as the seating history among judges to draw more insights on voting patterns.

2 Data

The original dataset contains opinions from 387,898 cases (1880-2013), collected by one of the authors, as well as features for these cases from "The United States Courts of Appeals database"

[8]. For this paper, we use a manually coded (or *labelled*) sample of 5% of all cases, where additional features cover the legal areas of the case, participants, and the motions involved. This data is randomly sampled among the years and weights are assigned to each circuit year according to the proportion of the universe of cases contained in the particular circuit and year. We also use a dataset of U.S. Courts of Appeals Judge biographies, from "The Judicial Research Initiative" [6].

3 Our Approach

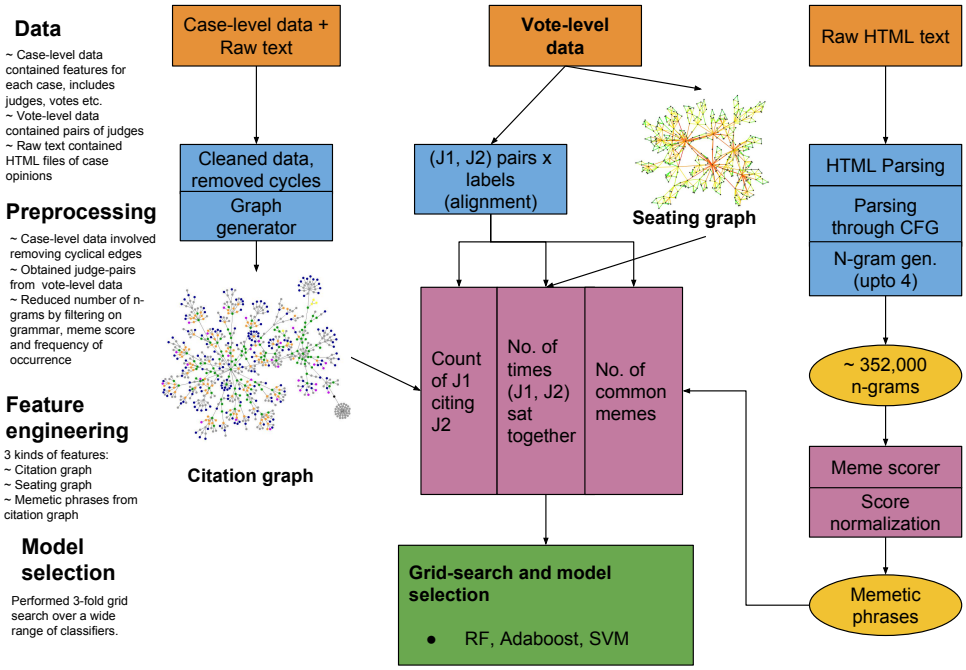

Figure 1: Data processing and machine learning pipeline.

We construct a number of features, belonging to the following main categories:

1. **Judge Bio**: We use data from The Judicial Research Initiative [6] and cross reference the judge's ID with the code for the judges in the case document to merge the two together. This gives us about 269 features [12]. Features included year of commission, law degree institution, years of service on a local court, etc.
2. **Case characteristics**: We use 228 features on case characteristics [11][13], history of the case, nature of the case, the participants and the issue coding. Features included year of decision, state of court, total number of appellants, type of the case, commonly used constitutional provisions etc.
3. **Proceedings of the case**: We use the text from the case document to extract out the case proceedings in the form of n-grams. We use the n-grams to generate memetic phrases. Commonly occurring n-grams and memes between judges were considered as features.
4. **N-grams, Citation and Seating patterns**: The seating and citation graphs provide data on how often two judges sat together and how often they cite each other. The raw opinion text was also used to generate n-grams, which were consequently labeled with a meme score. These were included as features for predicting vote alignment, and are discussed below.

3.1 Scoring memetic phrases

1. **Generating memes:** We generated n-grams of upto size 4, while filtering out n-grams that do not adhere to particular grammar rules. These grammar rules were chosen [9]

purposefully so that the resulting phrases conform to legal language, and were part of a context free grammar (CFG). These included:

2-grams: AN, NN, VN, VV, NV, VP.
3-grams: NNN, AAN, ANN, NAN, NPN, VAN, VNN, AVN,VVN, VPN,ANV,NVV,VDN, VVV, NNV, VVP,VAV,VVN, NCN,VCV, ACA, PAN.
4-grams: NCVN, ANNN, NNNN, NPNN, AANN, ANNN, ANPN, NNPN, NPAN, ACAN, NCNN, NNCN, ANCN, NCAN, PDAN, PNPN, VDNN, VDAN, VVDN.

2. **Meme score:** N-grams generated as per the CFG were scored on the basis of their *memeticity*. To quantify memes, we use the notion of memeticity defined in [10], which chiefly involves two factors: frequency and propagation. The frequency score of a phrase m is the ratio of cases that mention m in their opinion text to the total number of cases.

$$f_m = N_{\text{has meme}}/N_{\text{total}}$$

The propagation score measured the extent to which the cited phrase propagated over the citation graph,

$$P_m = \frac{d_{m\to m}}{d_{\to m}+\delta}/\frac{d_{m\to \not m}+\delta}{d_{\to \not m}+\delta}$$

where $d_{m\to m}$ = number of cases which cite m, and also cite at least one case which cites m; $d_{\to m}$ = number of cases which cite at least one case which cites m; $d_{m\to \not m}$ = number of cases which cite m, and do not cite any other case which cites m; $d_{\to \not m}$ = number of cases which do not cite any other case which cites m. δ is a noise factor to account for non-citing cases, and is taken to be 3.

The overall meme score of a phrase is therefore: $S_m = f_m \times P_m$.

3. **Scoring n-grams:** Using this definition of the meme score, we calculate the scores for each such n-gram in the 5% vote-level dataset. The score is generated by propagating along the topologically sorted set of nodes (opinions). This meme scorer algorithm is defined as:

SCORE-MEMES(N, NG, Adj)

```
▷ Iterate over all nodes in the citation network, N
for node ∈ N
    do
        ▷ Iterate over all n-grams in the node, N, and
        ▷ Using the n-gram dictionary, NG
        for gram ∈ NG[node]
            do
                ▷ Iterate over nodes in the citation network, with gram
                for other ∈ N, where gram ∈ NG[other]
                    do Update Meme Score
                        ▷ Process all adjacent nodes to other, N
                        for next ∈ Adj[node]
                            do Update Meme Score
                        ▷ O(E)
                ▷ O(V)
        ▷ O(N)
▷ O(V)
```

The complexity of this algorithm is $O(V^2NE)$, where V = number of vertices or cases, N = number of n-grams, E = number of edges or citations.

4. **Score normalization:** The meme score is finally normalized by the frequency of the meme across the network, so as to filter out non-memes such as *it is* or *have been*.
5. **Features:** We created two kinds of features - a) count of common memes b) count of common n-grams, between J1 and J2's opinions.

4 Experiments

We performed extensive grid search on a variety of models. Because the number of samples with the negative label (dissent) is very low (see Table 1), we use the label-averaged F1 score to evaluate models, and experimented with stratified sampling (SS) and class weighting (CW).

Label	Count	Percentage
Agree (+1)	106,947	95.9%
Disgree (-1)	4,591	4.1%

Table 1: Distribution of vote agreement and disagreement between judges.

After experimenting with a number of models and hyper-parameter tuning, we obtain the following results (Table 2):

Model	-1			+1			Avg.		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Baseline	0	0	0	0.96	1.0	0.98	0.46	0.49	0.47
Logistic	0.07	0.61	0.13	0.98	0.67	0.79	0.53	0.64	0.46
SVM, linear	0.04	0.97	0.07	0.98	0.05	0.09	0.051	0.52	0.28
SVM, Poly	0.04	0.97	0.07	0.98	0.05	0.09	0.51	0.51	0.08
Random Forests + SS	0.1	0.91	0.17	1.0	0.68	0.81	0.55	0.80	0.49
Random Forests + CW	0.34	0.47	0.39	0.97	0.99	0.98	0.66	0.73	0.69
AdaBoost + DT + CW	0.15	0.43	0.22	0.98	0.91	0.94	0.57	0.67	0.58
AdaBoost + RF + CW	0.48	0.48	0.48	0.98	0.98	0.98	0.73	0.73	**0.73**

Table 2: Results on training various models.

where the baseline is the majority classifier. Random Forests are training with stratified sampling (SS) and class weighting (CW), where the best class weighting was $\{+1 : 1, -1 : 25\}$. AdaBoost was used with decision trees (DT) and random forests (RF). The AdaBoost model with random forests and class weighting performed best.

5 Observations

We try to interpret the results of our models by listing down the most important features used by our best performing models.

5.1 What features play a major role in predicting the vote alignment?

The top 15 featuresof the best performing model, in order of decreasing importance, were:

1. Wlengthopin : Length of the judge's opinion
2. totalcites : Total number of citations in the opinion
3. votingvalence : Whether the voting is liberal or conservative or mixed
4. opinstat : Whether the opinion is identified by writer or per curiam
5. negativecites : Number of citations that are disapproving
6. decade2 : Time period of the case
7. day : Day of the case
8. common_n_grams : Common phrases (n-grams) used by the two judges
9. j2score : The second judge's historical percentage of agreement with majority (i.e., the non-writer signer's historical % of dissenting)
10. sat_together_count : The previous number of times that the pair of judges sat in the same panel

11. distance : The measure of difference between two judges' ideologies
12. state: The state where the case originated
13. treat: Treatment of decision below by appeals court (i.e., affirm, reverse, etc.)
14. liberalvote: Whether there is any vote on the case that can be categorized as liberal
15. month: Month in which the case occurred

We notice that the features 'common n-grams', and 'sat together count', which were generated generated from the judges' opinions and the seating graph respectively, were important. On the other hand, 'cite count', the number of times the judges cite one another, was not as important, and does not feature in this list.

To better understand these features, we classify them as "exogenous" and "endogenous", depending on whether they were determined by an external factor, such as the state or circuit, or an internal factor. We also use "network-based" to list important features that were engineered using the citation/seating/meme-networks (see Table 3).

Endogenous	**Exogenous**	**Network-based**
Wlengthopin	decade2	common n grams
totalcites	day	sat together count
opinstat	j2score	
votingvalence	distance	
negativecites	state	
liberalvote	treat	
	month	

Table 3: Important features in predicting vote alignment.

5.2 Memetic Phrases

As discussed, we generated memetic phrases using a Context-Free Grammar (CFG), pertaining to the possible legal phrases, and scored them by traversing the citation graph. We list some of the high-scoring meme phrases in Table 4.

Phrase	**Normalized Meme Score**
red heat	0.138
salvage services	0.0039
said cars	0.0029
Atlantic coast	0.00216
citizens of different states	0.00212
insurance effected	0.0020
separable controversy	0.0018
taken in tow	0.0017
schooner was	0.00126
fourteenth amendment	0.00125
contract of affreightment	0.00119
patented design	0.0011
constitution or laws	0.0009
mere transient or sojourner	0.0008

Table 4: Memes with the highest normalized meme scores across the citation network.

Upon observation, these phrases agree more with the definition of *memeticity*, and can be understood as legal phrases propagating over the citation network. For example, admirality law is a small area of law and its separation from other legal areas would tend to render phrases in admirality law cases to have high meme scores (cases that cite the meme-containing case are likely to themselves carry the meme and the number of progenitor cases that carry the meme are likely to be small). The memes that we generated were scored using the dictionary of n-grams from the entire 100% of citation graph, but span only 5% of the cases.

6 Discussion

We identified and tested a number of models to predict the vote alignment between judges on the U.S. Courts of Appeals, namely - Logistic Regression, Support Vector Machines, Random Forest, and Ensemble Methods like AdaBoost. We showed that these models significantly outperformed the baseline majority classifier on the averaged F1 score metric. As far as the authors are aware, these are first results when vote alignment between sitting judges on U.S. Courts of Appeals cases have been predicted.

Our work indicates that vote alignment between two judges can be predicted very accurately in the vast majority of cases. However, vote misalignment (or disagreement), is a harder problem, particularly due to the lack of labeled data. Since we performed these experiments on a 5% subset of the overall dataset, due to presence of hand-labeled features on cases present in this dataset, it is likely that the misalignment performance would improve with the entire dataset. Moreover, we found that features such as the number of times two judges sat together, and the number of common n-grams, were significantly important. From our results, this implies that judges who write opinions in a similar manner and sit together often are more likely to agree, while longer opinions, opinions with more citations in them, and the valence all contribute to determining when judges dissent.

References

[1] Katz, Daniel Martin and Bommarito, Michael James and Blackman, Josh, Predicting the Behavior of the Supreme Court of the United States: A General Approach (2014).

[2] Why The Best Supreme Court Predictor In The World Is Some Random Guy In Queens, Oliver Roeder, 2014.

[3] William Li, Pablo Azar, David Larochelle, Phil Hill, James Cox, Robert C. Berwick, Andrew W. Lo, Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions (2013)

[4] Harry Surden, Machine Learning and Law (2014).

[5] Law and Persuasion: The Language Behaviour of Lawyers, Walter Probert (1959).

[6] The U.S. Circuit Courts and the Federal Judiciary.

[7] Steven M. Beitzel, On Understanding and Classifying Web Queries (2006).

[8] Donald R. Songer, The United States courts of appeals database (SES-8912678).

[9] Elliott Ash, The political economy of tax laws in the U.S. states (Nov 2015).

[10] Tobias Kuhn, Matjaz Perc, and Dirk Helbing. Inheritance patterns in citation networks reveal scientific memes.

[11] Donald R. Songer. The United States Courts Of Appeals Data Base Documentation for Phase.

[12] Gerard S. Gryski, Gary Zuk. Multi-User Data Base on the Attributes of U.S. Appeals Court Judges, 1801-2000.

[13] Donald R. Songer. The United States Courts Of Appeals Data Base Documentation for Phase(Updated).